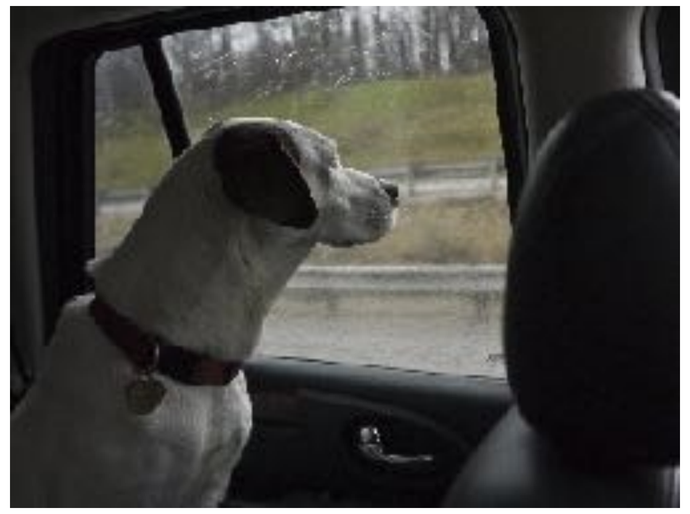
Vision Language Models
Image Captioning

Philipp Krähenbühl, UT Austin

Computer vision

"Traditional" Computer Vision
2014 - now

Task 1  Model 1 Output

Task 2  Model 2 Output

...

Unified Computer Vision
2020 - now

Task

Unified Model

Output

Image Classification

Applications

- Visual search
- Testbed for network design

Apple

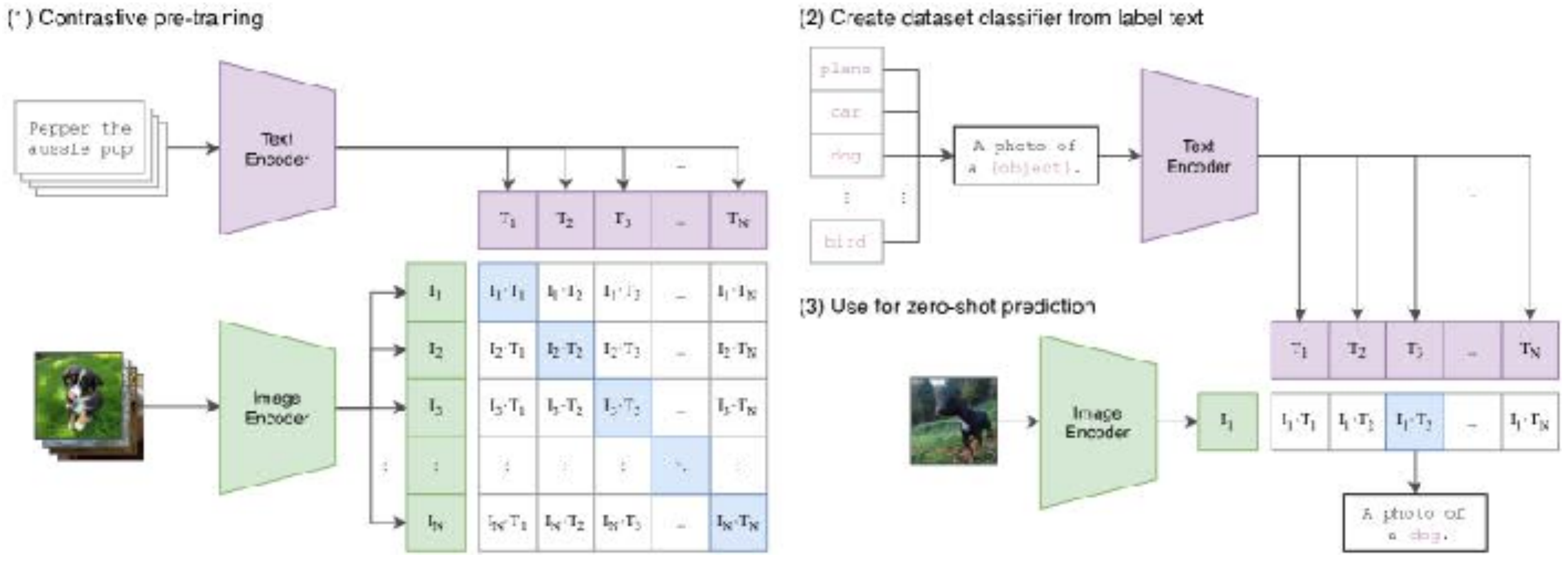
Image Captioning

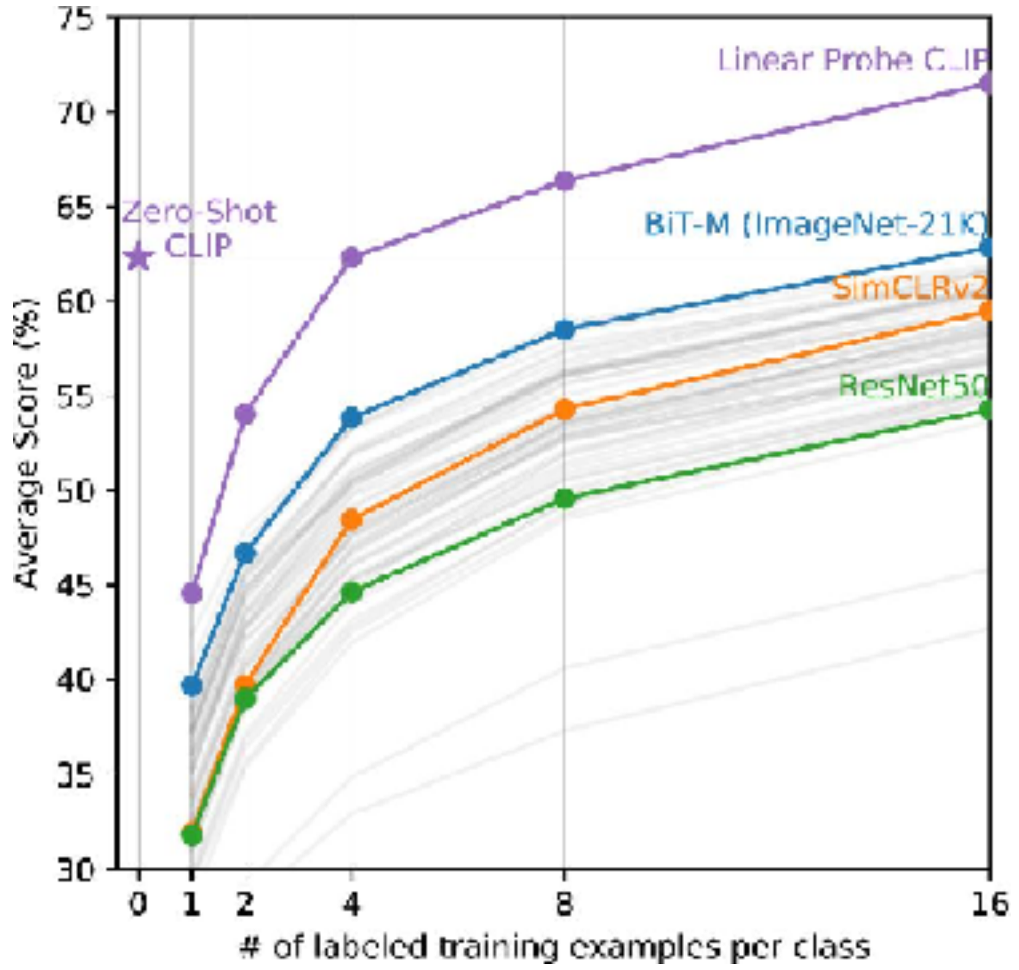
- Similar to classification
  - Richer annotation
  - Cheaper to obtain at scale
    - Alt-text on webpages

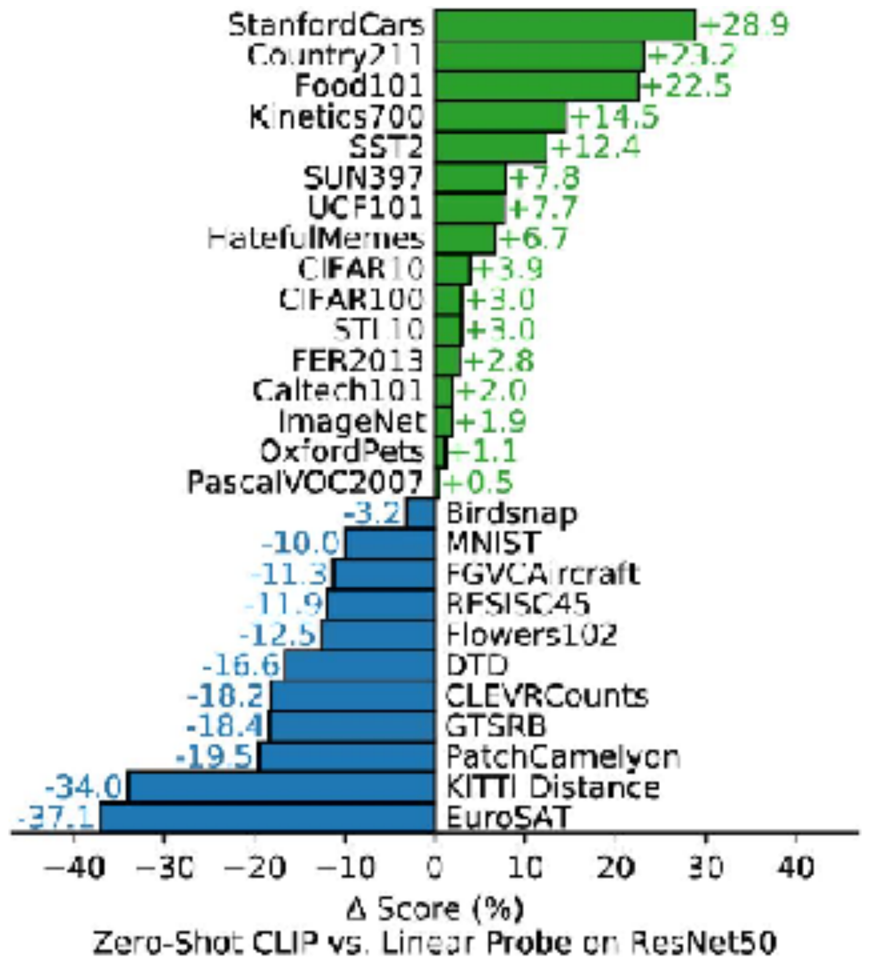
An image of an Apple cut into slices.

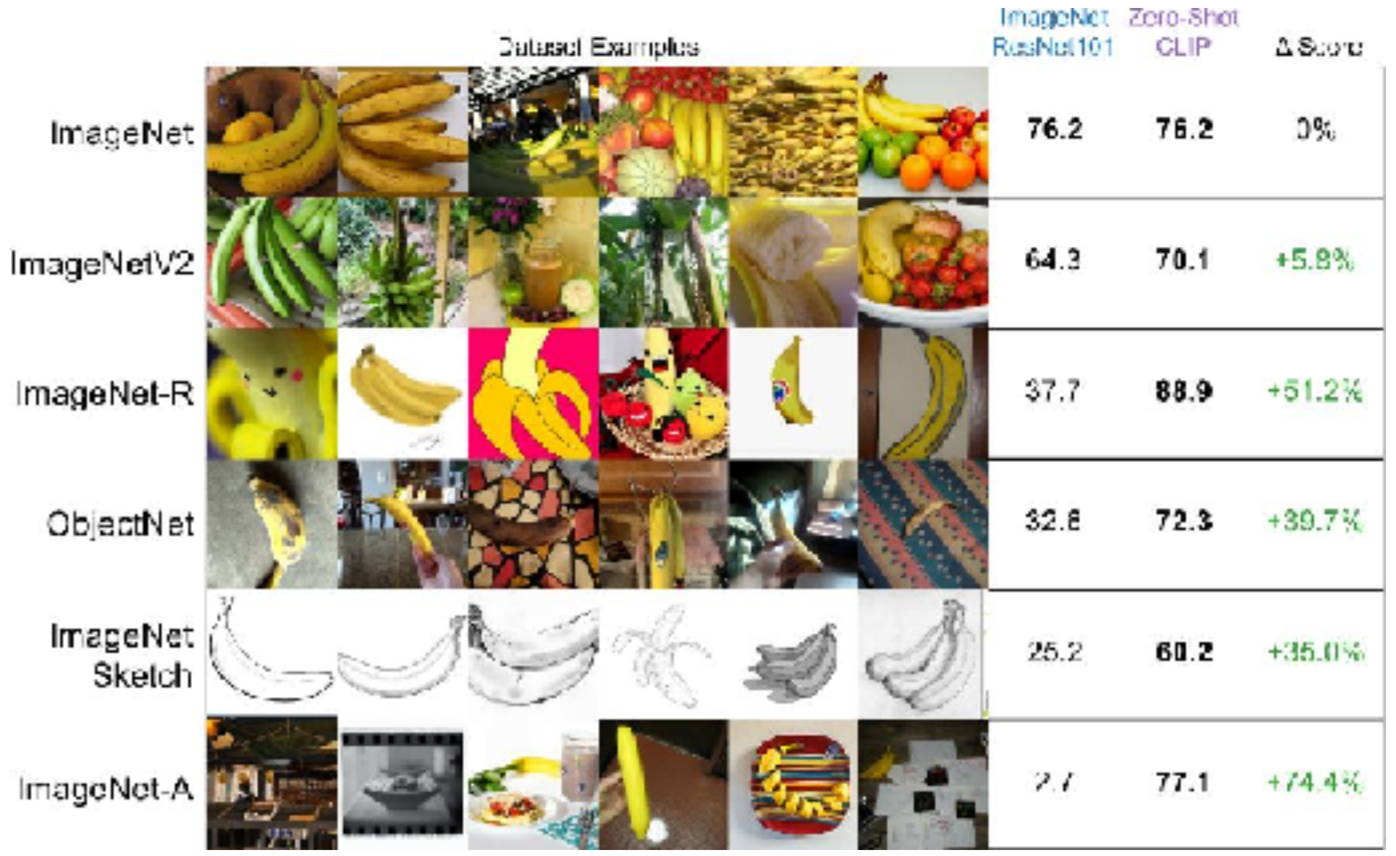
CLIP

- Download a large curated dataset of image and alt-text
- Learn an image + text embedding
    - With contrastive loss
- Great zero-shot classification performance



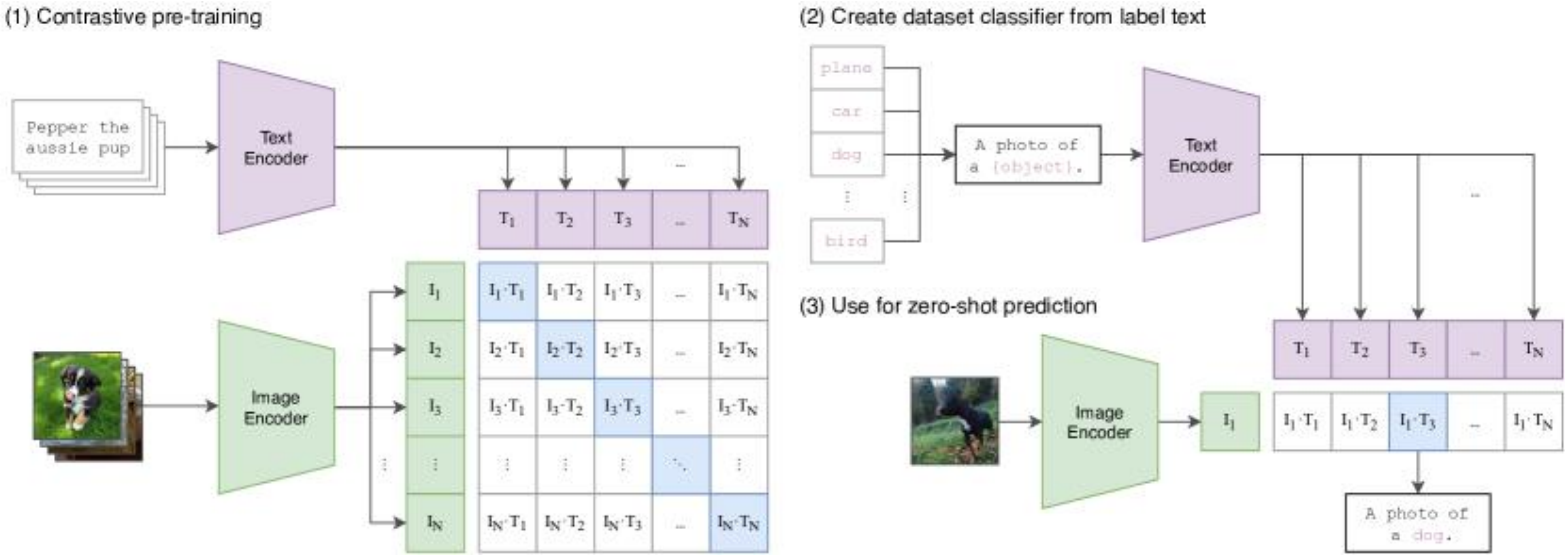






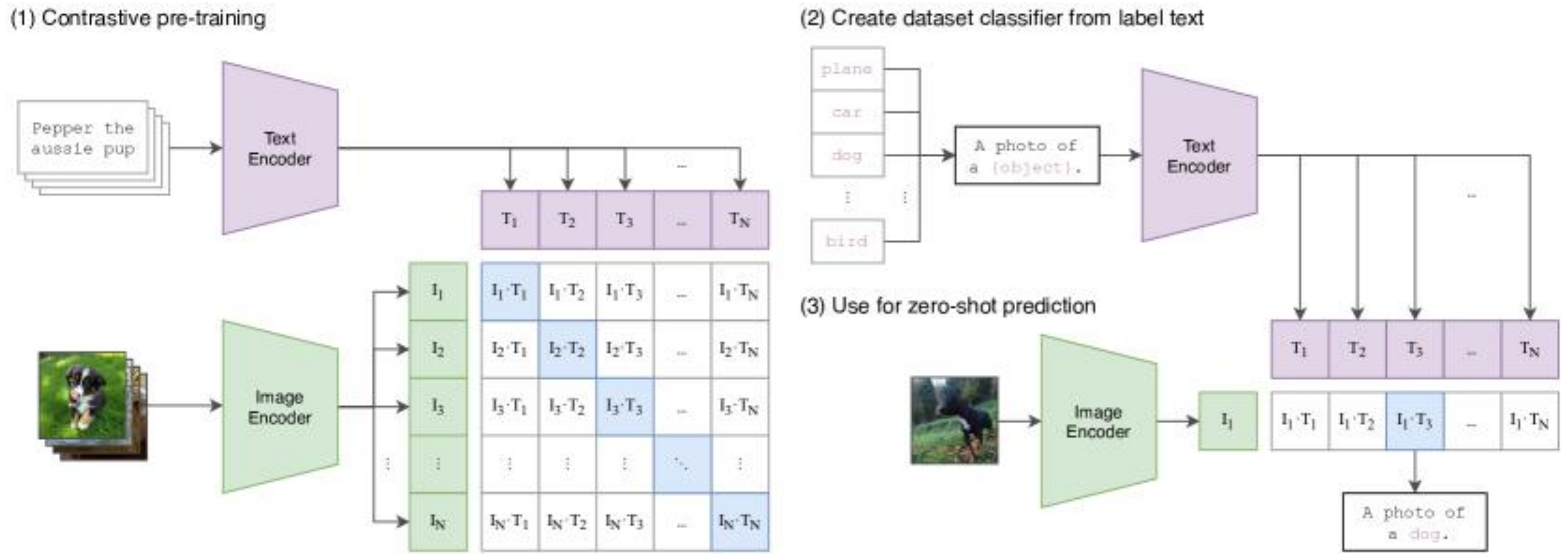
Learning Transferable Visual Models From Natural Language Supervision, Radford etal 2021

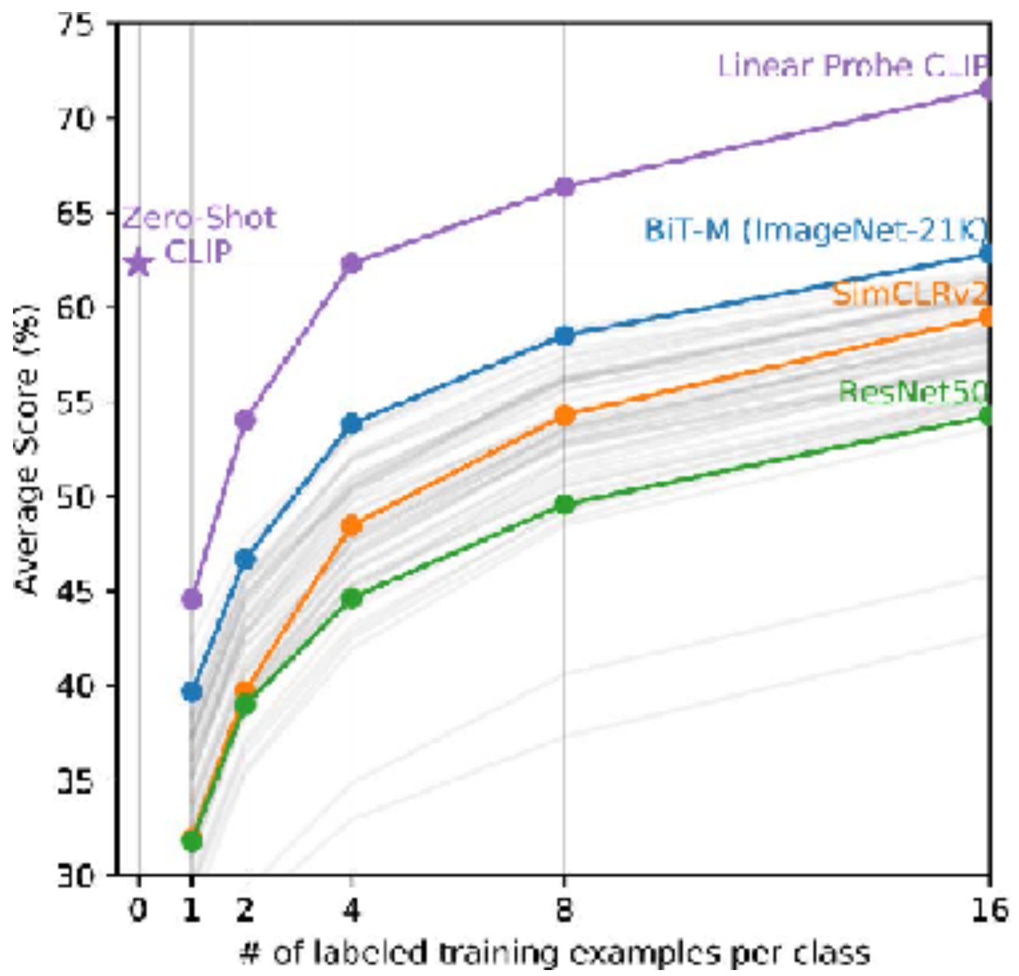
CLIP

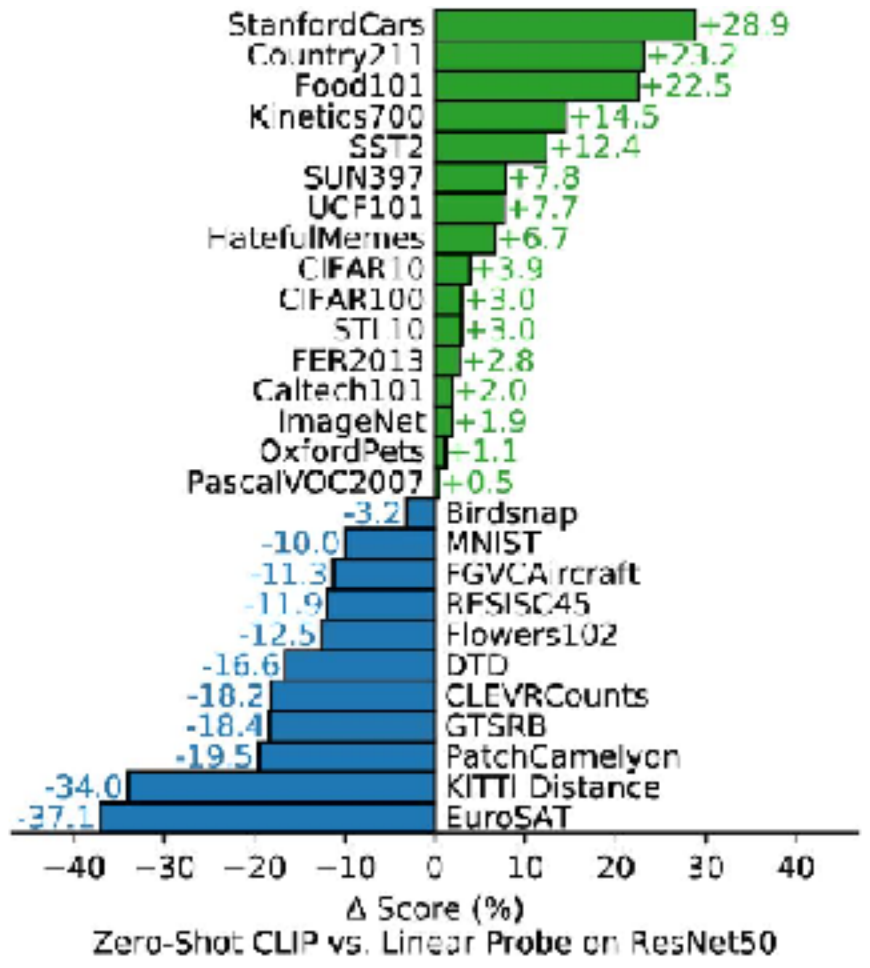


Learning Transferable Visual Models From Natural Language Supervision, Radford etal 2021

CLIP





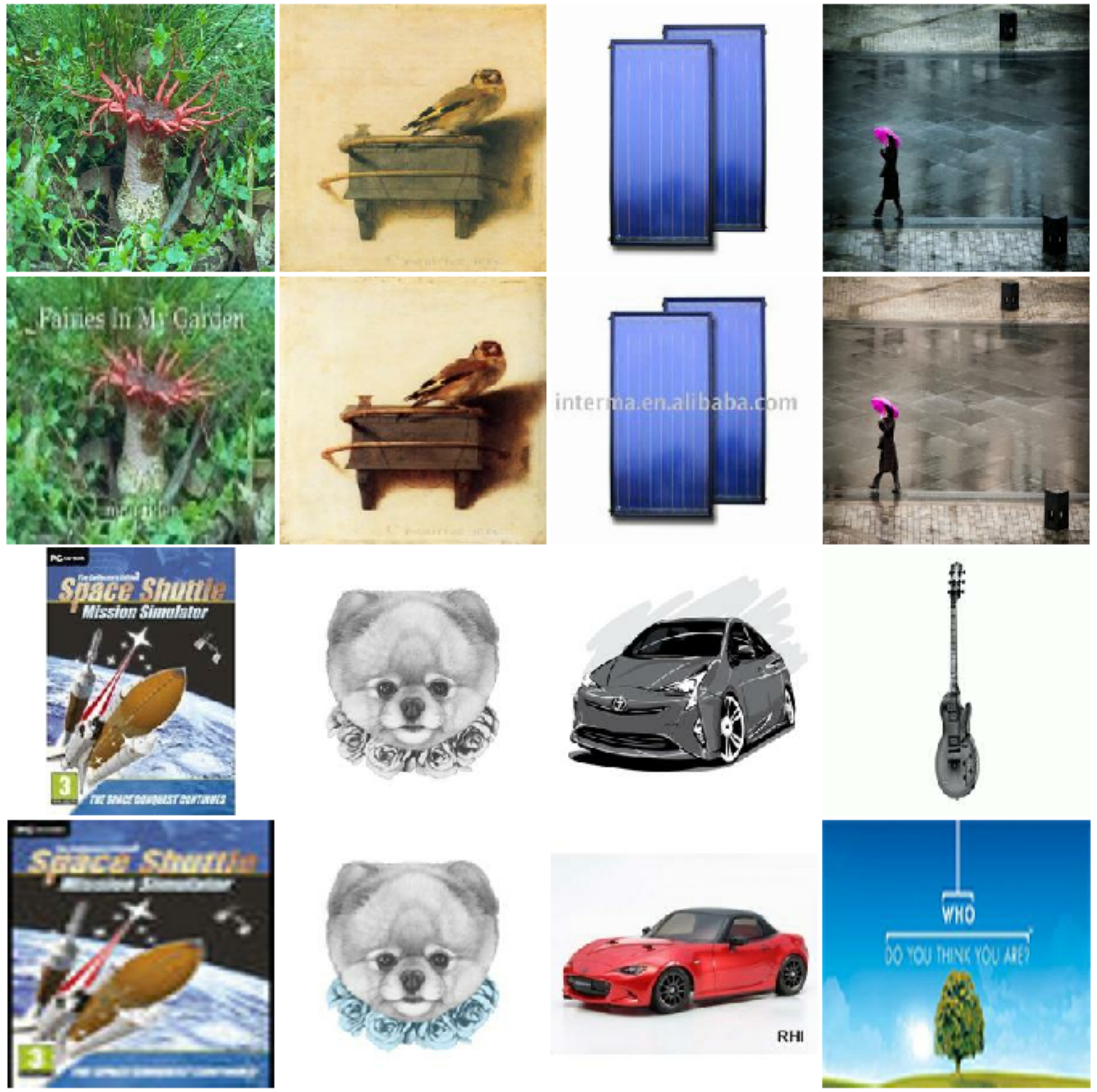


- Good classification model
- Loses many details of image
- Poor localization performance

Learning Transferable Visual Models From Natural Language Supervision, Radford etal 2021

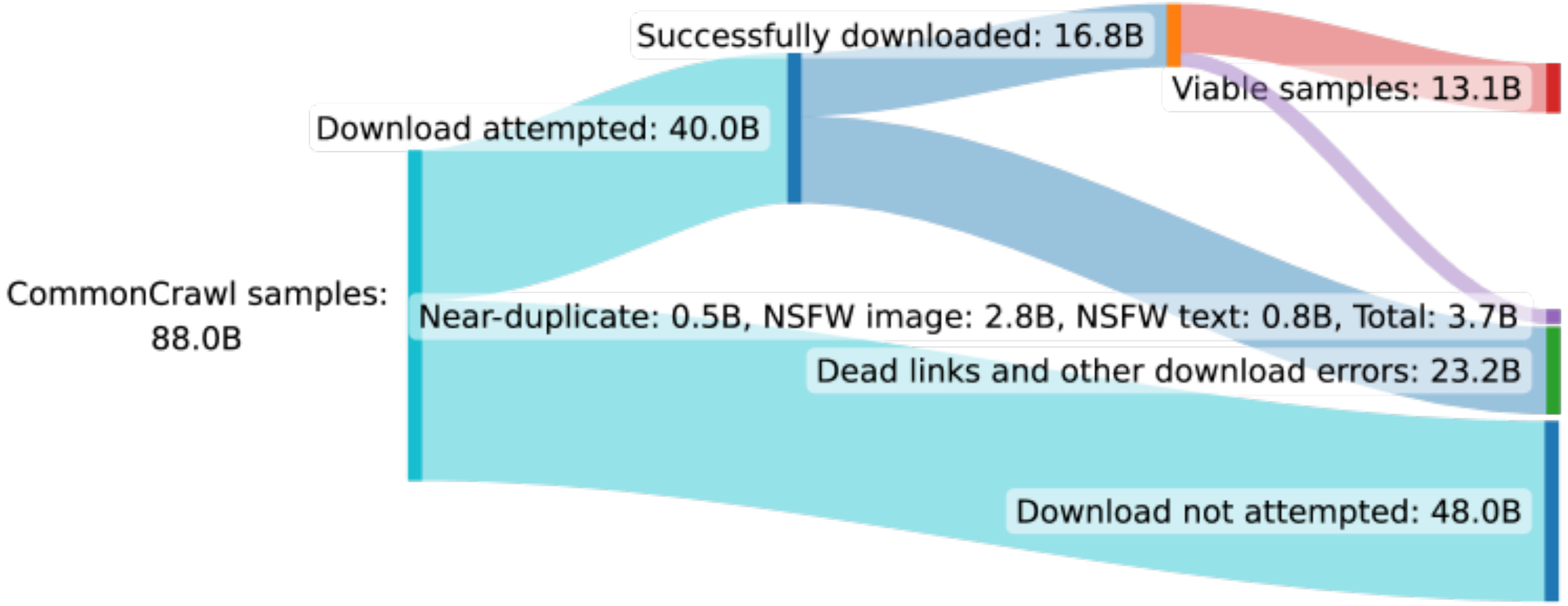
OpenCLIP / LAION

- New image-text dataset: LAION
  - Large scale (up to 2B)
  - DO NOT DOWNLOAD (NSFW)
- Starts to dig into data issues



Reproducible scaling laws for contrastive language-image learning, Cherti etal 2022

DataComp



- Largest image-text dataset yet: 13B images
  - Based on CommonCrawl
- Data filtering as a task
  - Fixed CLIP training
  - Standardized eval

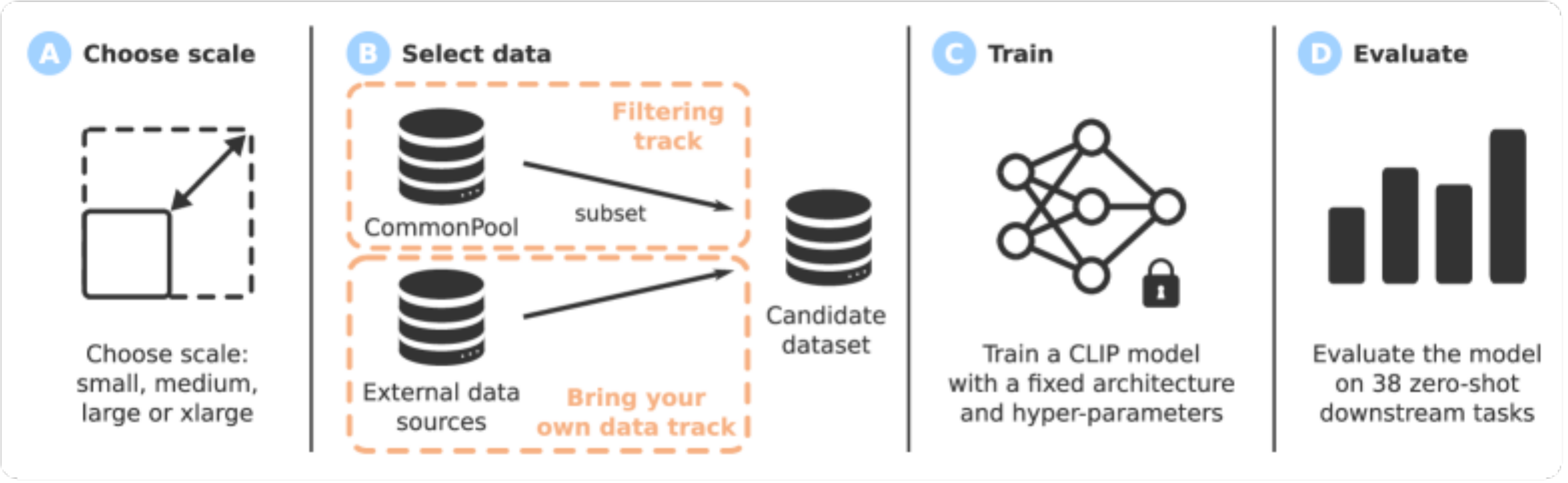


Table 1: Zero-shot performance of CLIP models trained on different datasets. DataComp-1B, assembled with a simple filtering procedure on image-text pairs from Common Crawl, leads to a model with higher accuracy than previous results while using the same number of multiply-accumulate operations (MACs) or less during training. See Section 3.5 for details on the evaluation datasets.

Dataset	Dataset size	# samples seen	Architecture	Train compute (MACs)	ImageNet accuracy
OpenAI's WIT [111]	0.4B	13B	ViT-L/14	1.1×10^{21}	75.5
LAION-400M [128, 28]	0.4B	13B	ViT-L/14	1.1×10^{21}	72.8
LAION-2B [129, 28]	2.3B	13B	ViT-L/14	1.1×10^{21}	73.1
LAION-2B [129, 28]	2.3B	34B	ViT-H/14	6.5×10^{21}	78.0
LAION-2B [129, 28]	2.3B	34B	ViT-g/14	9.9×10^{21}	78.5
DataComp-1B (ours)	1.4B	13B	ViT-L/14	1.1×10^{21}	**79.2**

DataComp: In search of the next generation of multimodal datasets, Gadre etal 2023

DataComp

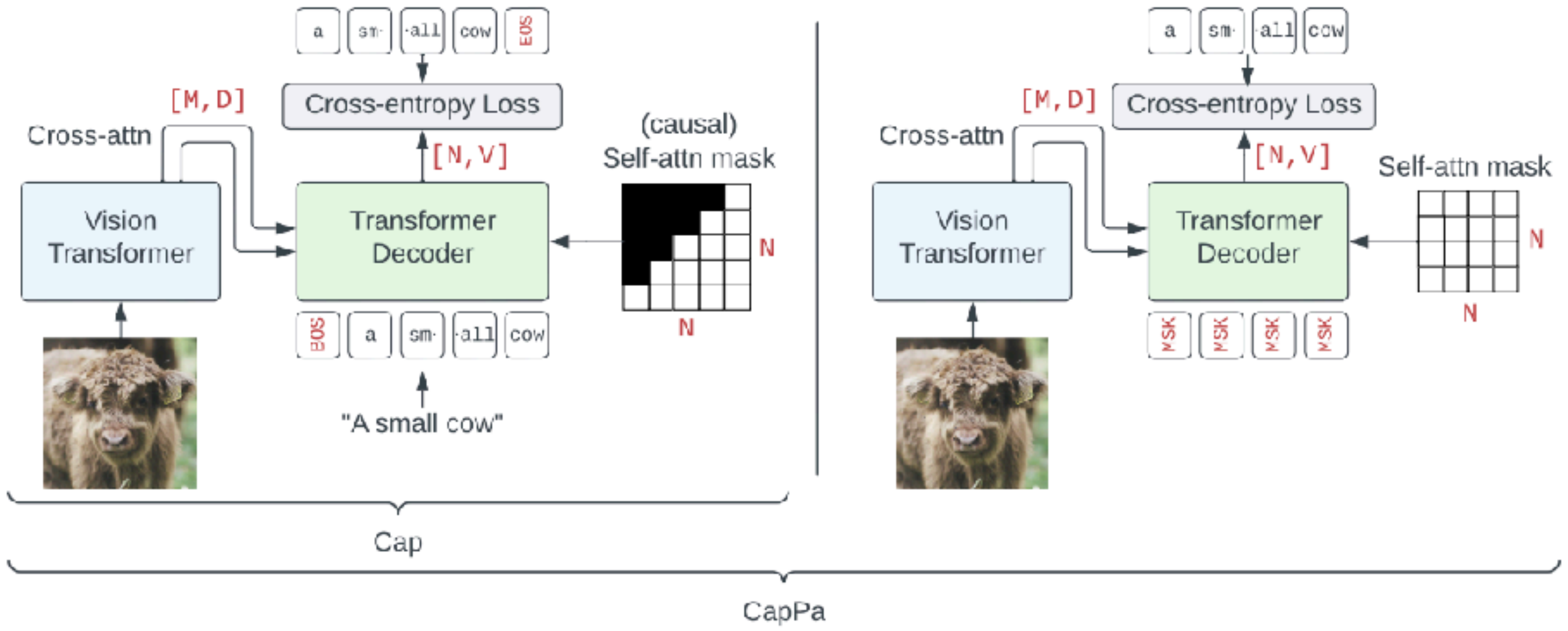
Scale	Filtering strategy	Dataset size	Samples seen	ImageNet	ImageNet dist. shifts	VTAB	Retrieval	Average over 38 datasets
small	No filtering	12.8M	12.8M	0.025	0.033	0.145	0.114	0.132
	Basic filtering	3M	12.8M	0.038	0.043	0.150	0.118	0.142
	Text-based	3.2M	12.8M	0.046	0.052	0.169	0.125	0.157
	Image-based	3M	12.8M	0.043	0.047	0.178	0.121	0.159
	LAION-2B filtering	1.3M	12.8M	0.031	0.040	0.136	0.092	0.133
	CLIP score (L/14 30%)	3.8M	12.8M	0.051	0.055	0.190	0.119	0.173
	Image-based ∩ CLIP score (L/14 30%)	1.4M	12.8M	0.039	0.045	0.162	0.094	0.144
medium	No filtering	128M	128M	0.176	0.152	0.259	0.219	0.258
	Basic filtering	30M	128M	0.226	0.193	0.284	0.251	0.285
	Text-based	31M	128M	0.255	0.215	0.328	0.249	0.307
	Image-based	29M	128M	0.268	0.213	0.319	0.256	0.312
	LAION-2B filtering	13M	128M	0.230	0.198	0.307	0.233	0.292
	CLIP score (L/14 30%)	38M	128M	0.273	0.230	0.338	0.251	0.328
	Image-based ∩ CLIP score (L/14 30%)	14M	128M	0.297	0.239	0.346	0.231	0.328
large	No filtering	1.28B	1.28B	0.459	0.378	0.426	0.419	0.437
	Basic filtering	298M	1.28B	0.516	0.423	0.446	0.480	0.458
	Text-based	317M	1.28B	0.561	0.465	0.465	0.352	0.466
	Image-based	293M	1.28B	0.572	0.454	0.483	0.479	0.476
	LAION-2B filtering	130M	1.28B	0.553	0.453	0.510	0.495	0.501
	CLIP score (L/14 30%)	384M	1.28B	0.578	0.474	0.538	0.466	0.529
	Image-based ∩ CLIP score (L/14 30%)	140M	1.28B	0.631	0.508	0.546	0.498	0.537
xlarge	No filtering	12.8B	12.8B	0.723	0.612	0.611	0.569	0.621
	LAION-2B filtering	1.3B	12.8B	0.755	0.637	0.624	0.620	0.636
	CLIP score (L/14 30%)	3.8B	12.8B	0.764	0.655	0.643	0.588	0.650
	Image-based ∩ CLIP score (L/14 30%)	1.4B	12.8B	0.792	0.679	0.652	0.608	0.663

- Very strong baselines
  - CLIP score filtering: discard images with low image-text similarity
  - Text-based filtering: fasttext and caption length filtering
  - Image-based filtering: Cluster clip-image encoder features (and filter according to distance to ImageNet)

Training data	Dataset size	# samples seen	ImageNet Acc.	Avg. performance (38 datasets)
OpenAI's WIT	0.4B	13B	75.5	0.61
LAION-400M	0.4B	13B	73.1	0.58
LAION-2B	2.3B	13B	73.1	0.59
LAION-2B	2.3B	34B	75.2	0.61
DataComp-1B	1.4B	13B	79.2	0.66

DataComp: In search of the next generation of multimodal datasets, Gadre etal 2023

CapPa

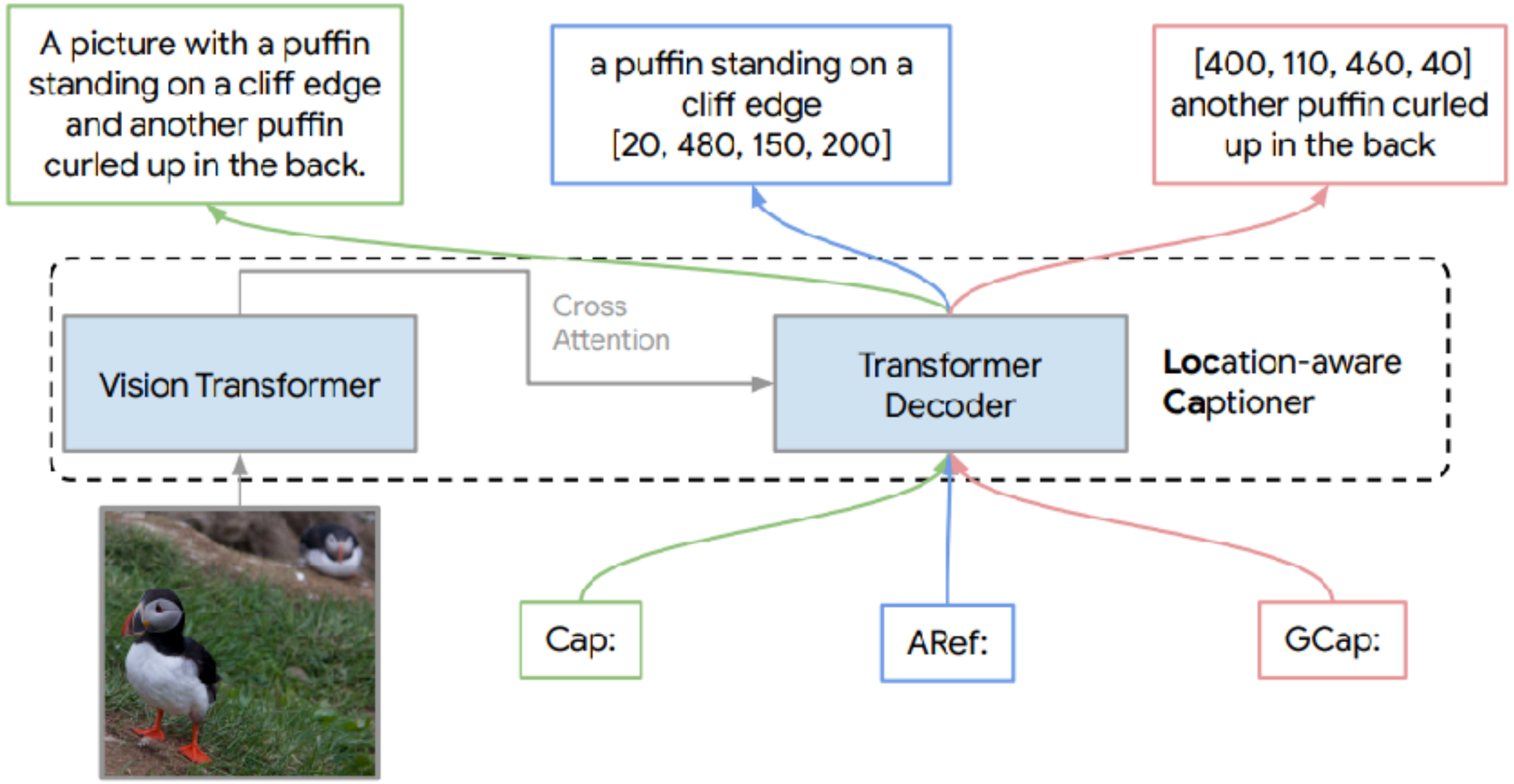


- Predict caption from image
  - Auto-regressively (easy for model)In parallel
  - (harder)
- Similar image classification performance than CLIP
- Better captioning and OCR performance
- Poor localization

Image Captioners Are Scalable Vision Learners Too, Tschannen etal 2023

LocCa

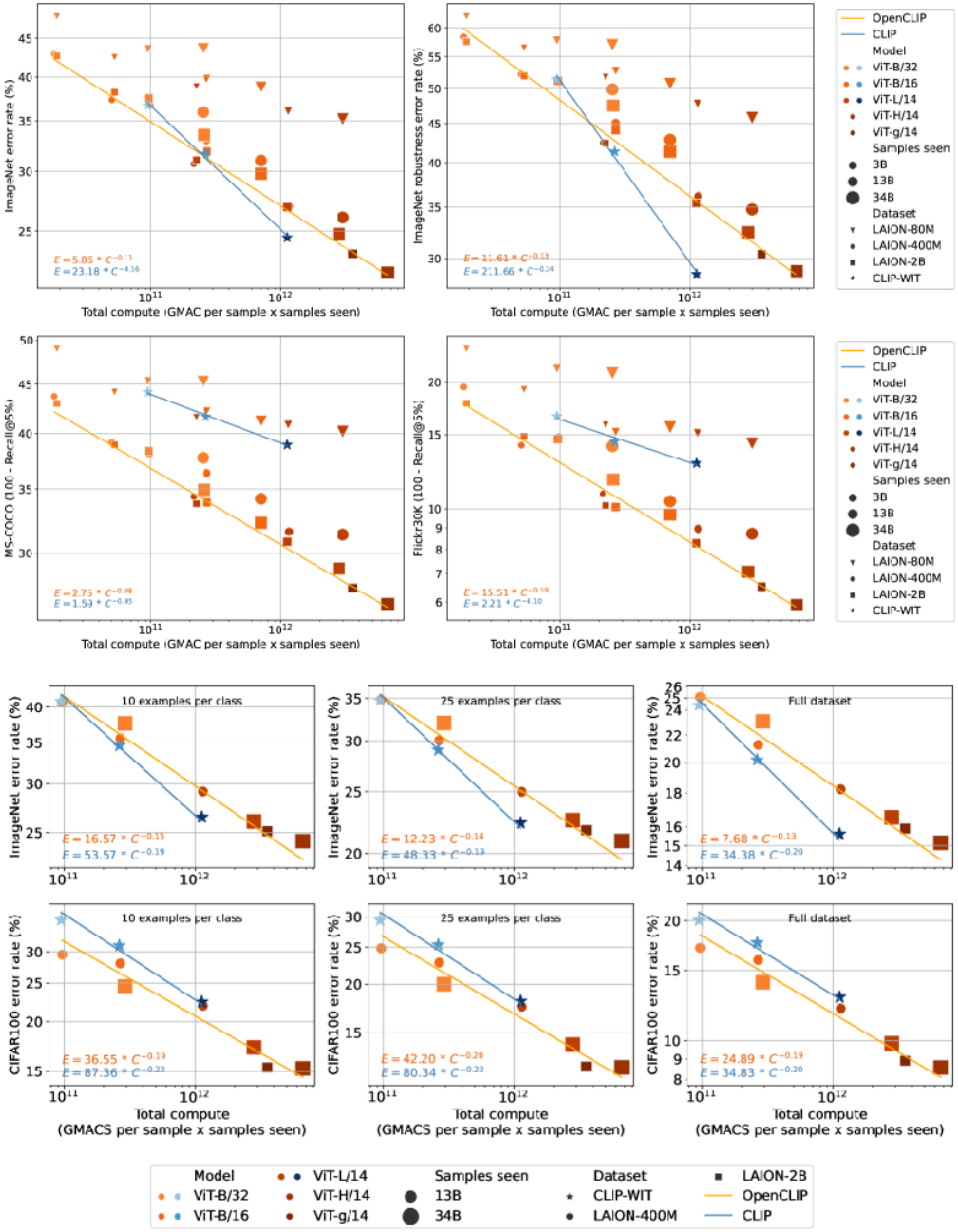
- Add location information to captions
  - Run an off-the-shelf open-vocabulary detector
- Tasks:
  - Captioning
  - Referring expression
  - Grounded captioning
- Similar classification performance
- Better detection, captioning, VQA



LocCa: Visual Pretraining with Location-aware Captioners, Wan etal 2024

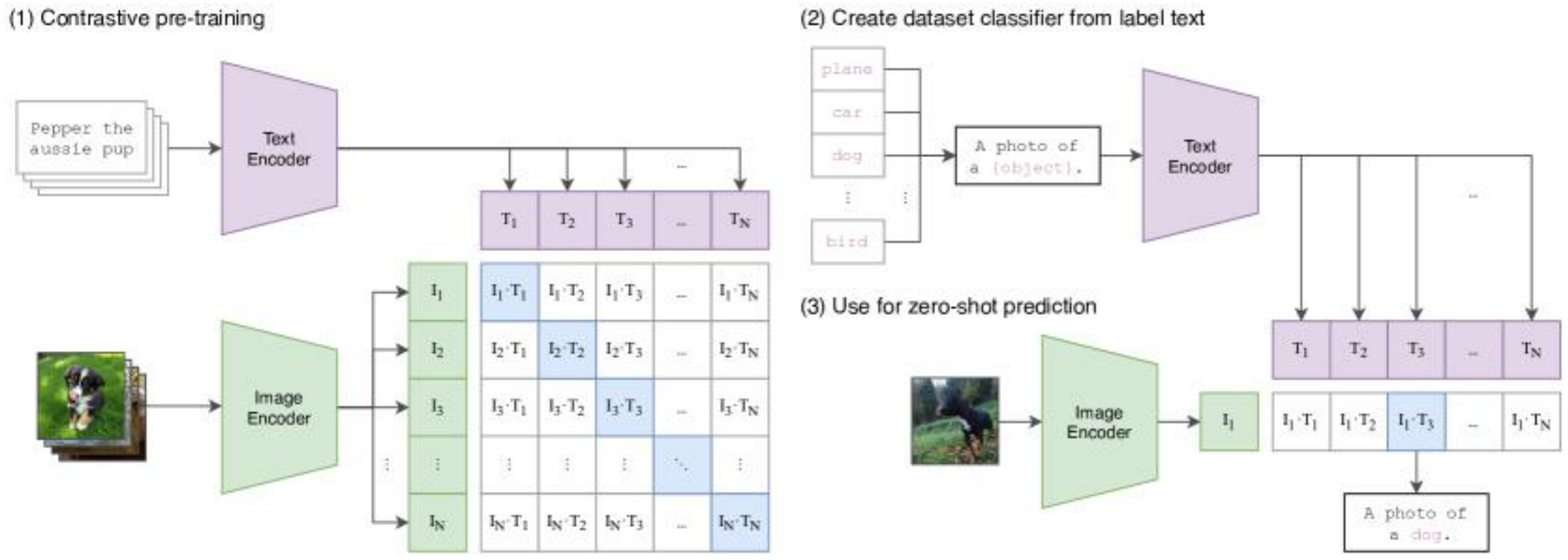
OpenCLIP / LAION

- Open Source replication of CLIP: OpenCLIP
  - Original CLIP may have overfit to ImageNet
- Highlights importance of data



Reproducible scaling laws for contrastive language-image learning, Cherti etal 2022

CLIP as a VLM

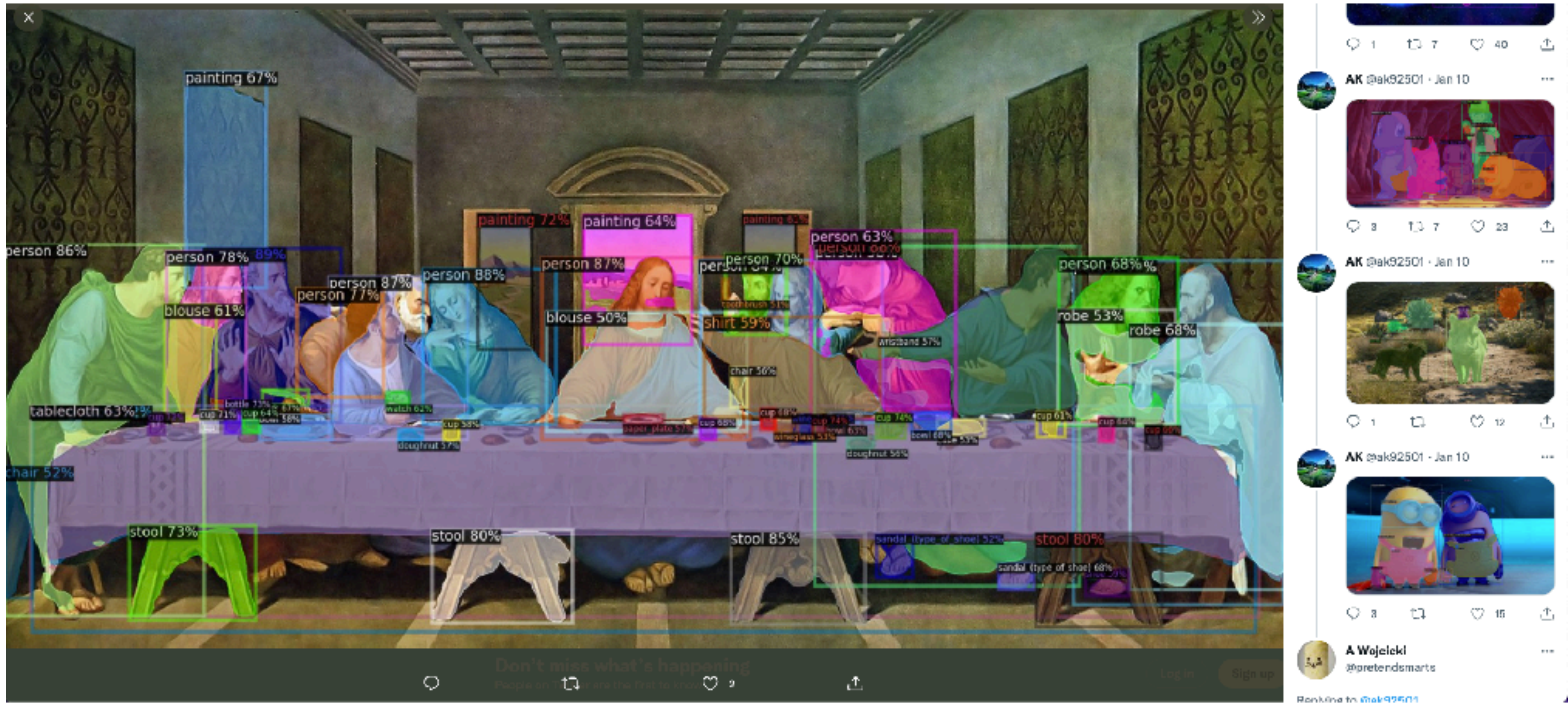


- Clip maps
  - Images to text
  - Text to images
- Primitive image and text models
- No dialogue

Image Captioning

Trend 1

- **Pre-training on caption datasets**
  - Much larger (size of internet)
  - Richer supervision



Open-Vocabulary Recognition

Trend 2

- **Image-encoders, text-encoders, text-decoders are stable**
  - Transformer
  - Form of ViT

Open-Vocabulary Recognition

Trend 3

- **Image data is exhausted, but information is not**
  - Datasets no longer grow significantly
  - What is annotated still grows

References

- Learning Transferable Visual Models From Natural Language Supervision, Radford etal. 2021
- Reproducible scaling laws for contrastive language-image learning, Cherti etal. 2022
- DataComp: In search of the next generation of multimodal datasets, Gadre etal. 2023
- Image Captioners Are Scalable Vision Learners Too, Tschannen etal. 2023
- LocCa: Visual Pretraining with Location-aware Captioners, Wan etal. 2024